様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　５月３１日現在

機関番号：１２６０８
研究種目：基盤研究（Ｂ）
研究期間：2010～2012
課題番号：22300050
研究課題名（和文）　ウィキペディアのモデル化に基づく解説型テキストの自動生成
研究課題名（英文）　Automatically Generating Expository Text by Modeling Wikipedia
研究代表者　藤井　敦（FUJII　ATSUSHI）
東京工業大学・大学院情報理工学研究科・准教授
研究者番号：30302433

研究成果の概要（和文）：
本研究は，様々な用語に関する説明を効率よく活用することを目的として，ウェブページの集合からウィキペディア記事のような解説型テキストを自動的に生成する手法について研究した．動物名や病名といった用語の種類によって説明に必要な観点が異なるため，ウィキペディアの記事集合から観点に基づく用語説明のパターンを学習する．用語の種類に応じて検索結果から必要な文章が抽出され，解説型テキストとして統合される．

研究成果の概要（英文）：
Aiming to efficiently utilize descriptions for various terms, we proposed a method to automatically generate expository text like an article in Wikipedia, from a set of pages on the Web. Because viewpoints required for describing a term are different depending on the type of that term, such as animal and disease, we use Wikipedia articles to learn viewpoint-based patterns for term descriptions. Depending on the type of a target term, we extract required sentences from a search result and integrate them into an expository text.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	3,200,000	960,000	4,160,000
2011 年度	6,500,000	1,950,000	8,450,000
2012 年度	4,100,000	1,230,000	5,330,000
年度			0
年度			0
総　計	13,800,000	4,140,000	17,940,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：自然言語処理

１．研究開始当初の背景

科学技術や文化の急速な発展によって，様々な用語について調べる機会が増えている．そのため，World Wide Web 上の様々なツールを使うことが多い．代表的なツールには，Google や Yahoo!などの検索エンジンとウィキペディアなどの人手で編集された事典がある．両者には，情報の量と質という点において，それぞれ長所と短所がある．

検索エンジンは，億単位のページ集合が検索の対象であり，提供される情報の量が多いという利点がある．しかし，検索される情報が体系化されておらず，必要のない情報も含まれるため，情報の質が低い．

事典は，説明を目的とした情報に限定され，項目等によって情報が統制されているため，情報の質が高いという利点がある．しかし，人手による編集に依存しているため，調べたい用語が必ず登録されているとは限らない．また，説明の内容が著者の視点に偏るという問題もある．すなわち，情報の量において問題がある．

２．研究の目的

本研究の目的は，検索エンジンと事典の長所を統合して，有用性が高い調べ物のツールを実現することである．既存の事典であるウィキペディアを分析することによって「用語説明が編集される仕組み」を解明し，用語説明に関するモデル（用語説明モデル）を構築する．さらに，そのモデルに基づいて検索エンジンの結果を組織化する．

用語説明モデルの構築において，「動物名」や「病名」といった対象によって説明に必要な観点が異なる点に着目した．例えば，「動物名」は「生態」や「形態」，「病名」は「症状」や「治療」といった観点に基づいて説明される．そこで，ウィキペディアから用語の種類に応じて異なる観点の構造を学習し，さらに観点ごとに固有の単語分布を学習する．その結果，例えば，動物の「ハクビシン」に関する検索の結果に含まれる複数のページやスニペットから，「生態」，「形態」，「分布」などの観点に対応するテキストを抽出し，「ハクビシン」に関する事典的な情報を組織化することを可能とする．

ここで，「ウィキペディアの存在が前提であれば，ウィキペディアの記事を読めばよいのではないか？」という疑問が生じるかもしれない．この問いに対する答えは「No」であり，本研究には 2 つの意義がある．まず，ウィキペディアの未登録語について，ウィキペディアと同じような観点の構造で説明を得ることができる．さらに，ウィキペディアの登録語に対しても，別の用語で使われている観点を補い，一般の Web ページから幅広く説明を収集することができる．

３．研究の方法

(1) 概要

本研究で提案する用語説明生成の概要を図 1 に示す．図 1 は，事前に行う「用語説明モデルの構築」と検索結果のテキスト集合が与えられた段階で実行する「検索結果の組織化」に大別される．用語説明モデルはウィキペディアの記事集合を用いて構築する．検索結果の組織化では，「りんご病」のような用語を検索質問として Web を検索した結果から，複数のテキスト（ページまたはスニペット）を収集し，用語説明モデルに基づいて観点ごとに個々のテキストを分類する．さらに，観点ごとに代表的なテキストをユーザに提示する．

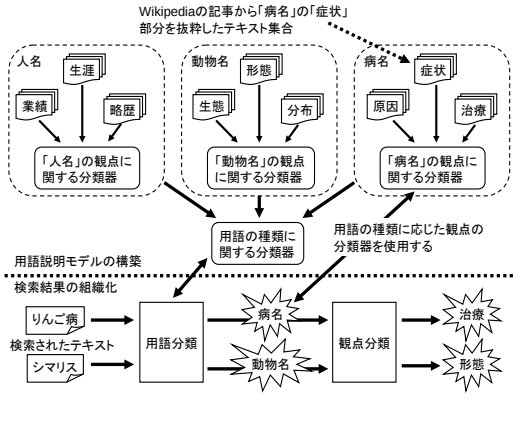

図 1　本研究の概要

(2) 用語説明モデルの構築

本研究で構築する用語説明モデルとは，「病名」や「動物名」といった用語の種類に応じて，説明の観点を列挙したプロトタイプである．さらに，ある用語について検索されたテキストが与えられると，用語の種類を特定し，その用語に対応する観点の候補から適切な観点を特定するための分類器である．

まず，「人名」，「動物名」，「病名」といった用語の種類ごとに，ウィキペディアの記事から観点の構造を抽出する．ここで問題となるのは，用語の種類をどのように特定するのかという点である．ウィキペディア記事には一つ以上のカテゴリが付与されているものの，「動物」のカテゴリには「脊椎動物」や「家畜」などに混ざって，「動物を題材とした作品」のように「動物」の体系に適さない記事も含まれている．そこで，ウィキペディアの記事集合をクラスタリングすることで，用語の種類に相当する「用語クラスタ」を自動的に特定する．

次に，ウィキペディアの記事にある「目次」

に着目する．例えば，「破傷風」に関するウィキペディアの記事を考えると，「目次」には「1.原因」，「2.症状」，「3.治療」などの項目（セクション）が並んでいる．ここでは，一つのセクションを一つの観点として使用する．しかし，あらゆる病名の記事でセクションが完全に一致するわけではない．そこで，ある用語の種類（例えば「病名」）ごとに複数の記事を収集し，使用頻度が高いセクションを観点として選択する．さらに，該当するウィキペディアの記事集合から複数の分類器を学習する．ただし，分類の目的によって使用する学習データが異なる．「用語の種類」を分類するためには，用語の種類ごとに記事集合をまとめて一つのカテゴリに対応する学習データを作る．他方で，「観点」を分類するためには，観点ごとに記事集合をまとめて一つのカテゴリに対応する学習データを作る．分類器の学習にはサポートベクターマシン（SVM）を使用した．

(3) 検索結果の組織化

検索結果の組織化では，「りんご病」などの用語について検索されたテキストを入力し，「用語分類」と「観点分類」を順番に実行する．用語分類では，用語に関する分類器を用いて，「りんご病」の説明テキストを「人名」，「動物名」，「病名」のいずれかに分類する．観点分類では，分類された用語の種類に応じて観点を分類する．図 1 の例では，「病名」に分類されたため，病名に対応する「原因」，「症状」，「治療」のいずれかに説明テキストを分類する．SVM に基づく One-Vs-Rest 法では分類結果ごとにスコアが計算されるため，観点ごとにスコアが高いテキストを抽出し，ユーザに提示する．

本研究における「検索結果の組織化」は，与えられた文字列を何らかの観点に分類するという抽象度の高い処理である．そこで，応用方法は 1 通りではなく，運用状況等に応じて使い分ける必要がある．例えば，一般的な Web 検索エンジンで検索された文書集合を観点に基づいて分類する応用方法がある．

４．研究成果

具体例を用いて説明する．Yahoo!で「ハクビシン」を入力して検索した上位 100 件のスニペットを分類した．本研究で作成した用語説明モデルには，「動物名」に対して，「生態」や「分布」など合計 7 種類の観点が対応していた．分類したスニペットには，上記 7 種類以外の観点として，「ハクビシン」をテーマにした「著作」や「名称の由来」があった．これらの用語説明モデルに含まれない観点は，検索されたスニペットの内容を分析して抽出するしかない．今後は，カテゴリ分類とクラスタリングの併用について検討する必要がある．

多義語の例として，「キーウィ」で検索されたスニペットを分類した．用語分類において，鳥のキーウィについて記述されたスニペットは「動物名」に分類され，果物のキーウィについて記述されたスニペットは「植物名」に分類された．さらに，これらのスニペットは観点分類によって動物名や植物名に固有の観点に分類された．本手法によって多義や多観点を考慮した検索を実現できることが分かった．

さらに「用語説明モデルの構築」に焦点を当てて実験を行った．実験に使用したデータは約 5000 件のウィキペディア記事であり，各記事は人手によって以下の 10 種類に分類されている．

動物，映画，病気，企業，人物，植物，虫，料理，魚類，スポーツ

表 1 は，ウィキペディア記事のクラスタリングによって特定された用語クラスタと，各用語クラスタに対応付けされた観点の一覧である．表 1 の「用語クラスタ」を見ると，「動物」が「動物 1」と「動物 2」に分割されており，その代わりに「魚類」がなくなっている．しかし，表 1 の「観点」を見ると，用語クラスタごとに人間の直感に合う観点の名称が抽出されていることが分かる．

表 1：自動構築された用語クラスタと観点

クラスタ	観点
虫	特徴，分類，生態，形態
スポーツ	歴史，ルール
映画	キャスト，スタッフ，ストーリー，あらすじ，登場人物
病気	治療，症状，原因，検査，診断，分類，疫学，予後，病態，歴史，予防
動物 1	特徴，歴史
料理	歴史，作り方
企業	沿革，事業所，主な商品，関連会社，歴史，会社概要，主な製品
人物	経歴，略歴，人物，来歴・人物，著書，来歴，生涯，生い立ち，パーソナル，エピソード
植物	特徴，利用，分類，主な種
動物 2	生態，形態，人間との関係，分類，分布，特徴，亜種，利用，近縁種

５．主な発表論文等

〔雑誌論文〕（計 7 件）

① Atsushi Fujii, Yuya Fujii, and Takenobu Tokunaga. Effects of Document Clustering in Modeling Wikipedia-style Term Descriptions. Proc. of the 8th International Conference on Language Resources and Evaluation, pp.2543-2546, 2012. 査読有

② Odbayar Chimeddorj and Atsushi Fujii. Enhancing Lemmatization for Mongolian and its Application to Statistical Machine Translation. Proc. of the 10th Workshop on Asian Language Resources, 2012. 査読有

③ Odbayar Chimeddorj and Atsushi Fujii. Enhancing Lemmatization for Mongolian Using Part-of-Speech Information. Proc. of the 5th International Universal Communication Symposium, 2011. 査読有

④ 藤井 敦. アンカーテキストモデルと検索質問分類によるWeb文書検索の高度化. 情報処理学会論文誌, Vol.51, No. 12, pp.2330-2342, 2010. 査読有

⑤ Atsushi Fujii, Seiji Takegata. Question Answering for the Operation of Software Applications: A Document Retrieval Approach. IEICE Transactions on Information and Systems, Vol.E93-D, No.6, pp.1369-1377, 2010. 査読有

⑥ Atsushi Fujii. Modeling Wikipedia Articles to Enhance Encyclopedic Search. Proceedings of the 7th International Conference on Language Resources and Evaluation, pp.2591-2595, 2010. 査読有

⑦ Shihono Karikome and Atsushi Fujii. A System for Supporting Dietary Habits: Planning Menus and Visualizing Nutritional Intake Balance. Proc. of the 4th International Conference on Ubiquitous Information Management and Communication, pp.386-391, 2010. 査読有

〔学会発表〕（計 9 件）

① 中山 祐輝, 藤井 敦. レビューテキストを対象とした評価条件の抽出手法. 言語処理学会第 19 回年次大会発表論文集, pp.248-251, 2013.03.14, 名古屋大学（愛知）.

② 吉成 祐人, 藤井 敦. 造語の過程に基づく派生オノマトペの抽出. 言語処理学会第 19 回年次大会発表論文集, pp.366-369, 2013.03.14, 名古屋大学（愛知）.

③ 伊藤 玄暉, 椛沢 直樹, 藤井 敦. 直喩と比較による間接表現を利用した用語説明の自動生成. 言語処理学会第 19 回年次大会発表論文集, pp.596-599, 2013.03.15, 名古屋大学（愛知）.

④ 西原 弘真, 苅米 志帆乃, 藤井 敦. 料理レシピを対象としたアウトライン型自動要約. 情報処理学会 第 89 回デジタル・ドキュメント研究会 第 110 回情報基礎とアクセス研究会合同研究会, 2013.02.28, 東洋大学（東京）.

⑤ 苅米 志帆乃, 藤井 敦. 料理レシピテキストを対象とした構造解析の高精度化. 電子情報通信学会 データ工学研究会, 2013.06.06, 国立情報学研究所（東京）.

⑥ 藤井 裕也, 藤井 敦, 徳永 健伸. Wikipedia記事構造のモデル化による用語説明の自動編集. 言語処理学会第 18 回年次大会発表論文集, pp.1059-1062, 2012.03.16, 広島市立大学（広島）.

⑦ 岡田 瑞穂, 藤井 敦. レビューテキスト間の類似度を用いた協調フィルタリング. 言語処理学会第 18 回年次大会発表論文集, pp.711-714, 2012.03.15, 広島市立大学（広島）.

⑧ 中島 正貴, 藤井 敦. 造語の過程に基づく複合オノマトペの検出手法. 言語処理学会第 18 回年次大会発表論文集, pp.69-72, 2012.03.14, 広島市立大学(広島）.

⑨ 佐々木 智, 藤井 敦. 回答の根拠を提示する意思決定支援型の質問応答システム. 言語処理学会第 17 回年次大会発表論文集, pp.252-255, 2011.03.18, 豊橋技術科学大学（愛知）.

６．研究組織

(1)研究代表者

藤井　敦（FUJII　ATSUSHI）

東京工業大学・大学院情報理工学研究科・准教授

研究者番号：30302433

(2)研究分担者

徳永　健伸（TOKUNAGA　TAKENOBU）

東京工業大学・情報理工学（系）研究科・教授

研究者番号：20197875